Appendix 1: detail of training ROI auto-segmentation

Data split: to enhance the training of our automatic segmentation model, we allocated as many samples as possible to the training set, reserving 20 samples for validation of the segmentation algorithm. In the validation set, we applied an early stopping strategy after 32 epochs to preserve the optimal model parameters.

Pre-processing: the pre-processing phase commenced with the standardization of the anatomical orientation of both images and labels according to the RAS (right, anterior, superior) axis codes. We then resampled the images and labels to a uniform voxel spacing of 1 mm × 1 mm × 1 mm using bilinear interpolation for images and nearest neighbor interpolation for labels. Furthermore, we adjusted the intensity values of the images to a standardized range of (–1,000, 150) Hounsfield units through linear transformation, incorporating an optional clipping step to manage outliers. The final step involved removing the background by isolating the foreground region in both images and labels, utilizing a mask derived from the original image.

Training process

Data augmentation: during the training phase, sub-volumes were selectively cropped from the images and labels to balance the representation of positive and negative labels. This involved defining specific spatial dimensions and sample quantities. To enhance the diversity of the training dataset, online data augmentation techniques such as adjustments in spacing and random cropping were employed, ensuring a varied image set for each training iteration.

Loss function: for the loss function, we adopted the DiceCELoss, integrating the Dice Loss and Cross-Entropy Loss. This hybrid approach combines the benefits of both loss functions to effectively manage the challenges associated with class imbalance and segmentation accuracy.

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\left(\alpha \log\left(\frac{2}{\frac{1}{\in + Dice_i} + \frac{1}{\in + CE_i}}\right) + (1-\alpha)\log\left(\in + Dice_i\right)\right) \quad [1]$$

Here, N represents the batch size, $Dice_i$ and CE_i denote the values of the Dice Loss and Cross-Entropy Loss functions for the i^{th} sample, while α serves as a weighting parameter for balancing the influence of both loss functions. By assigning zero weights to unlabeled pixels, the learning process can focus exclusively on the labeled ones, allowing for generalization across the entire volume.

Hyper parameters: we employed the Adam optimizer with an initial learning rate of 1e–3, 300 epochs, with early stopping after 32 rounds.

https://dx.doi.org/10.21037/tlcr-24-995

Table S1 The definition of radiological characteristics in the study

Radiological features	Description
Lobulation sign	The surface of the nodule is uneven, like the lobes of a leaf
Spiculation sign	Several linear shadows of varying lengths on the margins of the nodule extending into the surrounding lung tissue
Vacuole sign	Air-containing hypodense areas of 1–3 mm within the lesion
Air bronchogram sign	Inflated bronchial tubes cross the interior or margins of the lesion
Vascular convergence sign	Stenosis and occlusion of branch vessels after penetration into the interior of the nodule
Pleural retraction sign	The linear or curtain-shaped dense shadow formed by adhesion and pulling of a pulmonary nodule to the adjacent pleura
Density	It is expressed by the mean HU of the largest cross-section at the level of the tumor. Divide it into three segments as ordinal date: the density <–400 HU, low density; –400 HU ≤ the density <–100 HU, medium density; the density ≥–100 HU, high density
CTR	Ratio of the maximum diameter of the solid component of the nodule to the maximum diameter of the nodule. It is divided into four segments as ordinal date: CTR <10%, low; 10%≤ CTR <50%, sub-low; 50%≤ CTR <90%, sub-high; CTR ≥90%, high
Max diameter	The longest diameter of the tumor in the largest cross-section

CTR, consolidation/tumor ratio; HU, Hounsfield unit.

Table S2 The novel IASLC grading system of invasive lung adenocarcinoma

Grade	Differentiation	Patterns
1	Well-differentiated	Lepidic predominant with no or less than 20% of HGPs (solid, micropapillary, and complex glandular patterns)
2	Moderately differentiated	Acinar or papillary predominant with no or less than 20% of HGPs
3	Poorly differentiated	Any tumor with 20% or more of HGPs

HGP, high-grade pattern; IASLC, International Association for the Study of Lung Cancer.

Table S3 Univariate and multivariable analysis of clinical features and radiological characteristics

Characteristics	Univariable		Multivariable	
	OR (95% CI)	P value	OR (95% CI)	P value
Gender (male)	2.55 (1.69, 3.87)	<0.001		0.073
Age (years)	1.01 (0.99, 1.03)	0.248		
Smoking	3.32 (2.00, 5.52)	<0.001		0.879
Spiculation	3.61 (2.07, 6.30)	<0.001		0.264
Lobulation	2.02 (1.08, 3.78)	0.029		0.444
Vascular convergence	1.69 (0.53, 5.43)	0.375		
Pleural retraction	3.19 (1.96, 5.20)	<0.001		0.533
Bronchogram	0.74 (0.48, 1.14)	0.172		
Vacuole	1.29 (0.82, 2.03)	0.271		
Density	6.36 (4.26, 9.47)	<0.001	2.27 (1.23, 4.21)	0.009
Max diameter	1.95 (1.36, 2.80)	<0.001		0.272
CTR	4.49 (3.27, 6.17)	<0.001	2.50 (1.54, 4.04)	<0.001

Explanations of radiological features are detailed in the supplementary material. A P value <0.05 indicates a significant difference. CTR, consolidation/tumor diameter ratio; OR, odds ratio; CI, confidence interval.

https://dx.doi.org/10.21037/tlcr-24-995

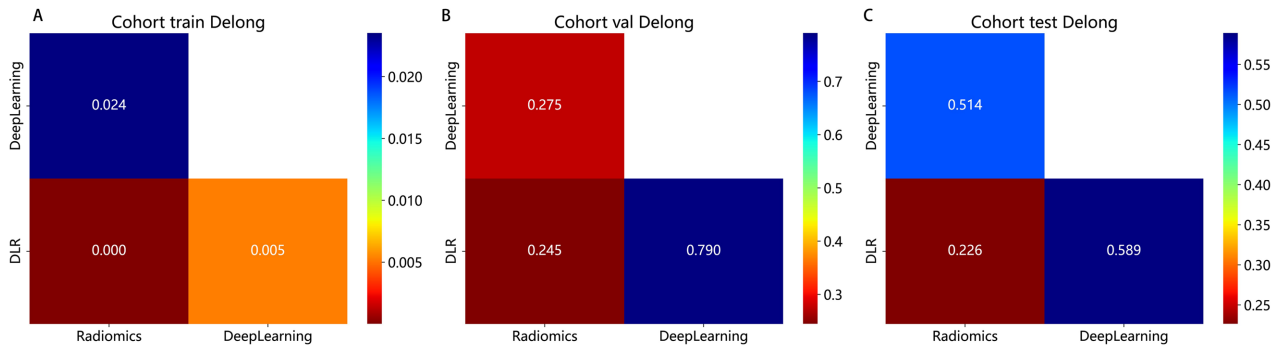

Figure S1 DeLong' test between models in each cohort. DLR, deep learning radiomics.

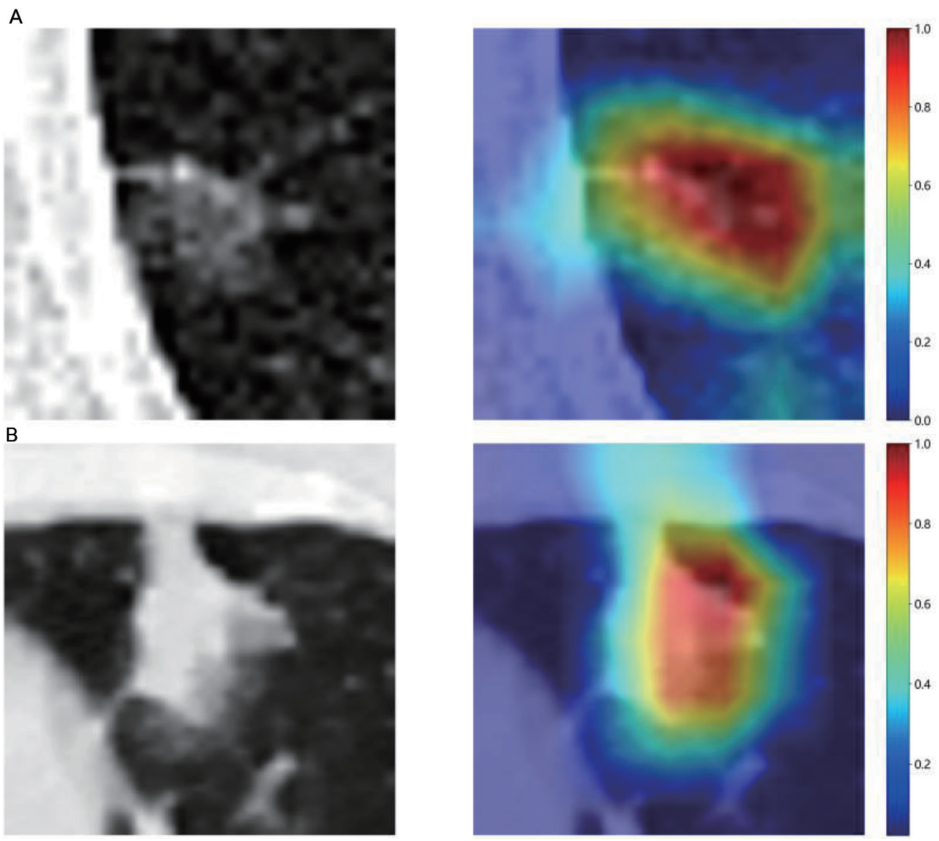

Figure S2 The Grad-CAM visualizations for two samples. These visualizations are instrumental in demonstrating how the model focuses on different regions of the images to make its predictions. Grad-CAM, gradient-weighted class activation mapping.

https://dx.doi.org/10.21037/tlcr-24-995

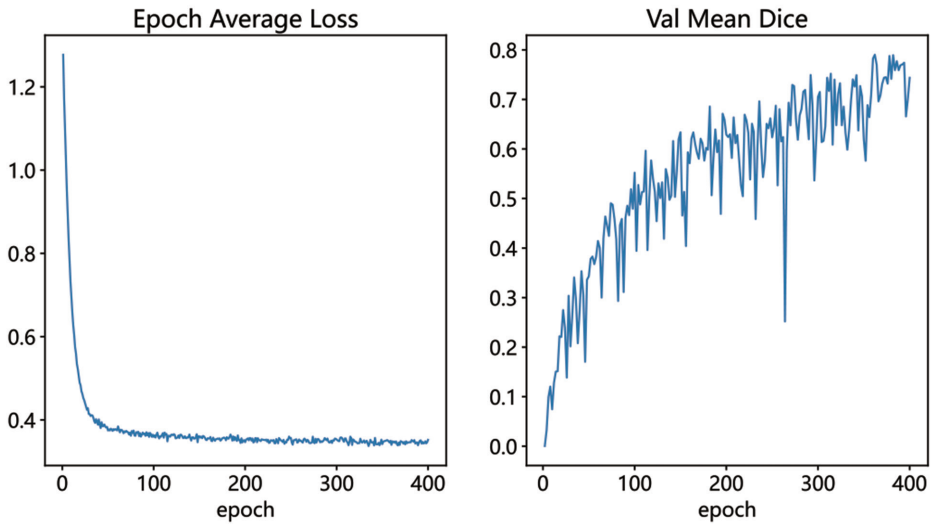

Figure S3 Training process of VNet.

Table S4 Dice of each model in training and validation cohorts

Cohort	UNet	SegResNet	VNet
Training	0.626	0.563	0.900
Validation	0.605	0.572	0.836

Evaluation metrics

In the prediction of ROI regions, we utilize a sliding window approach with dimensions of 48×48×48 voxels to process the input data. Throughout this prediction phase, scattered points—erroneously identified voxels—can occur throughout the image space. Given the characteristic continuous presence of ROI regions within the data space, we opt for the largest connected ROI area (utilizing the KeepLargestConnectedComponent function in MONAI) as our final prediction outcome.

For assessing the segmentation accuracy, we employ the Dice similarity coefficient (DICE). The DICE coefficient is a widely used metric for measuring the overlap between two samples, providing an indication of their similarity. It is calculated as the size of the intersection divided by the average size of the two samples.

$$Dice(GT, Pred) = \frac{2Area_{GT} \bigcap Area_{Pred}}{Area_{GT} + Area_{Pred}} \quad [2]$$

During the model evaluation phase, we apply post-processing to select the largest connected ROI region, thereby enhancing our model's performance compared to the training phase's mean DICE scores. The accompanying figure (*Figure S3*) illustrates the changes in loss and DICE metrics throughout the VNet training process. *Table S4* presents the model's DICE performance following post-processing adjustments.

https://dx.doi.org/10.21037/tlcr-24-995

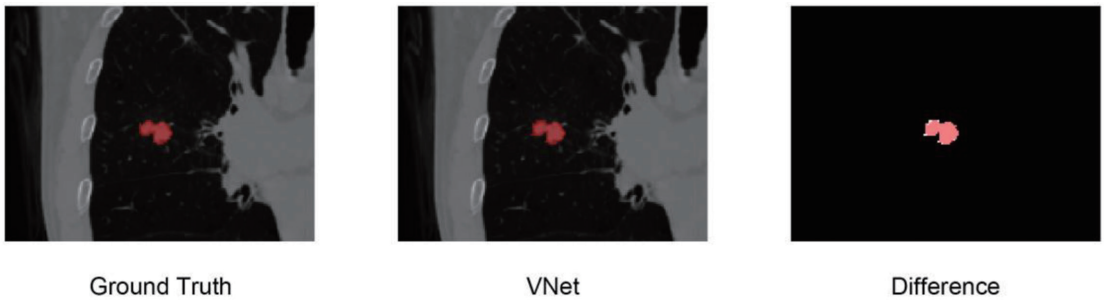

Figure S4 Visualization of results.

Visualization of results

The following *Figure S4* presents the recognition results of our VNet. In the rightmost "Diff" section, it can be observed that the differences in recognition accuracy are minimal. Such discrepancies are considered negligible within our deep learning workflow. This effectively validates the feasibility of the automatic delineation process we have proposed.

Appendix 2: details of deep learning model training

Data preparation

Crop ROI: in our methodology, for each patient, we selected the slice that presented the largest ROI as the representative image. To reduce complexity and minimize background noise in our algorithmic analysis, we retained only the smallest bounding rectangle encompassing the ROI. This rectangle was expanded by an additional 10 pixels.

Data augmentation: our approach involved standardizing the intensity distribution across RGB channels through Z-score normalization of the images. These normalized images were then utilized as inputs for our networks. During the training phase, we implemented real-time data augmentation strategies, including random cropping, horizontal flipping, and vertical flipping. For test images, we restricted processing to normalization only.

Model training

Transfer learning: in this study, we explored the performance of prominent networks such as VGG19, incpetion_v3, ResNet50, ResNet101, and DenseNet121 to enhance the performance of traditional CNN-based models. Additionally, we conducted comparative analyses of these networks to identify the most suitable algorithm for our specific research requirements.

Hyper parameters: in our study, to ensure the model's effectiveness across various patient populations with notable variability, we implemented transfer learning. This process involved initializing the model with pre-trained weights from the ImageNet database, enhancing its adaptability to diverse datasets. A critical aspect of our approach was the meticulous adjustment of the learning rate to foster better generalization across datasets. For this purpose, we employed the cosine decay learning rate strategy, defined as follows:

$$\eta_t = \eta_{\min}^i + \frac{1}{2}\left(\eta_{\max}^i - \eta_{\min}^i\right)\left(1+\cos\left(\frac{T_{cur}}{T_i}\pi\right)\right) \quad [3]$$

The notation $\eta_{\min}^i$ =0 sets the minimum learning rate, while $\eta_{\max}^i$ =0.01 establishes the maximum learning rate. The term T_i =20 denotes the number of epochs in the iterative training process. Other essential hyperparameters include the use of stochastic gradient descent (SGD) as the optimizer and softmax cross-entropy for the loss function.

https://dx.doi.org/10.21037/tlcr-24-995

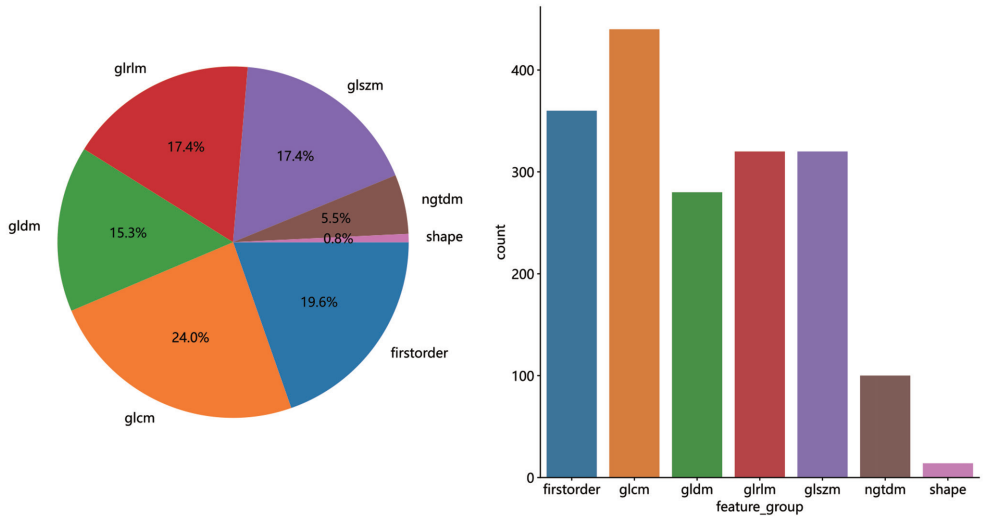

Figure S5 Number and ratio of handcrafted features.

Appendix 3: handcrafted feature extraction

In our research, we organized manually engineered radiomic features into three main categories: (I) geometry; (II) intensity; and (III) texture. The geometry category is focused on quantifying the three-dimensional geometrical attributes of the tumor. The intensity category is concerned with evaluating the statistical distribution of voxel intensities within the tumor using first-order statistics. On the other hand, the texture category examines the patterns and spatial arrangements of voxel intensities through more complex second-order and higher-level analyses. For the extraction of texture features, we employed various methodologies, such as the gray-level co-occurrence matrix (GLCM), gray-level run length matrix (GLRLM), gray-level size zone matrix (GLSZM), and the neighborhood gray-tone difference matrix (NGTDM). The extraction process for each subregion was performed using the pyradiomics tool (version 3.0.1), in compliance with the standards set by the Imaging Biomarker Standardization Initiative (IBSI).

Statistics on handcrafted features: in this investigation, we extracted a total of 1836 handcrafted radiomic features, which were divided into three key groups: shape, first-order statistics, and texture, comprising 360 first-order features, 14 shape features, and a wide range of texture features, respectively. These features were extracted through a proprietary tool developed using Pyradiomics, detailed at http://pyradiomics.readthedocs.io. The allocation of these manually engineered features into their respective categories is graphically represented in *Figure S5*.

Appendix 4: radiomics model

LASSO-based radiomics feature selection: we implemented LassoCV. This approach, coupled with 10-fold cross-validation, was used for the selection of radiomics features. The details of this process are visually represented in *Figure S6*.

Metrics: although the AUC of the XGBoost classifier is second only to the LightGBM classifier, the accuracy is relatively higher. The XGBoost classifier demonstrated a notable AUC of 0.917 in the training cohort with a 95% CI of 0.880–0.954. In the validation cohort, the XGBoost classifier achieved an AUC of 0.772 (95% CI: 0.666–0.878), outperforming the LR, RandomForest, ExtraTrees, and MLP classifiers in terms of AUC. Moreover, in the test cohort, XGBoost recorded an AUC of 0.771 (95% CI: 0.671–0.871). The details are shown in *Table S5* and *Figure S7*.

https://dx.doi.org/10.21037/tlcr-24-995

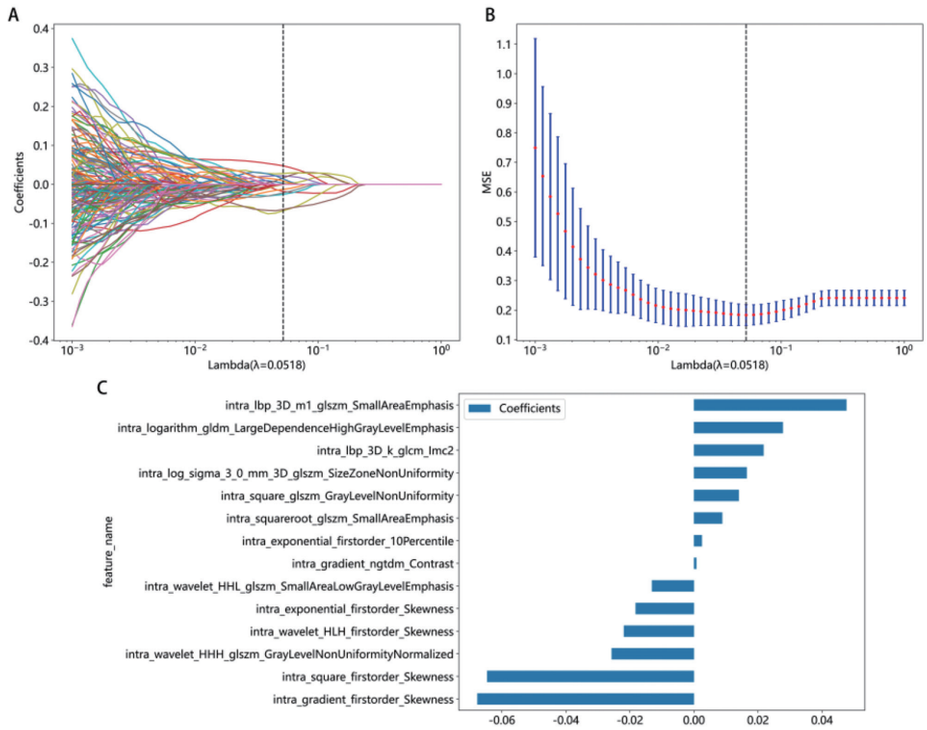

Figure S6 The coefficients derived from the LASSO during 10-fold cross-validation, as applied in radiomics model.

Table S5 Radiomics model results

Classifier name	Cohort	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV
LR	Training	0.812	0.881	0.835–0.928	0.818	0.807	0.735	0.872
	Validation	0.771	0.731	0.615–0.847	0.548	0.877	0.680	0.803
	Test	0.602	0.702	0.599–0.804	0.865	0.443	0.485	0.844
RandomForest	Training	0.897	0.959	0.937–0.982	0.943	0.867	0.822	0.959
	Validation	0.760	0.772	0.669–0.876	0.645	0.815	0.625	0.828
	Test	0.765	0.793	0.697–0.890	0.730	0.787	0.675	0.828
ExtraTrees	Training	0.749	0.796	0.737–0.855	0.727	0.763	0.667	0.811
	Validation	0.719	0.768	0.666–0.870	0.677	0.738	0.553	0.828
	Test	0.776	0.796	0.701–0.890	0.649	0.852	0.727	0.800
XGBoost	Training	0.857	0.917	0.880–0.954	0.875	0.844	0.786	0.912
	Validation	0.781	0.772	0.666–0.878	0.645	0.846	0.667	0.833
	Test	0.735	0.771	0.671–0.871	0.649	0.787	0.649	0.787
LightGBM	Training	0.812	0.886	0.842–0.930	0.875	0.770	0.713	0.904
	Validation	0.688	0.800	0.701–0.899	0.871	0.600	0.509	0.907
	Test	0.653	0.787	0.691–0.883	0.081	1.000	1.000	0.642
MLP	Training	0.857	0.917	0.879–0.955	0.807	0.889	0.826	0.876
	Validation	0.750	0.747	0.630–0.863	0.613	0.815	0.613	0.815
	Test	0.622	0.705	0.603–0.807	0.892	0.459	0.500	0.875

AUC, area under curve; CI, confidence interval; DLR, deep learning radiomics; NPV, negative predictive value; PPV, positive predictive value.

https://dx.doi.org/10.21037/tlcr-24-995

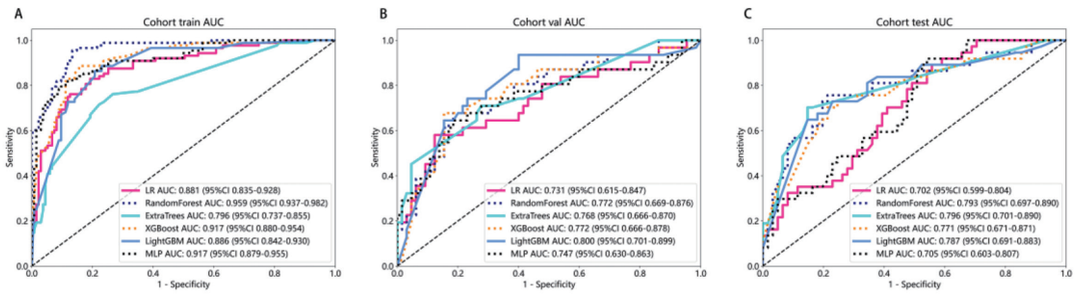

Figure S7 The performance of the classifiers on Rad model training, validation, and test datasets.

Table S6 Metric results for deep learning model

Network name	Cohort	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV
DenseNet121	Training	0.816	0.872	0.826–0.919	0.636	0.933	0.862	0.797
	Validation	0.802	0.848	0.759–0.937	0.742	0.831	0.676	0.871
	Test	0.653	0.703	0.595–0.810	0.757	0.590	0.528	0.800
Inception_v3	Training	0.812	0.838	0.782–0.895	0.602	0.948	0.883	0.785
	Validation	0.729	0.823	0.737–0.908	0.774	0.708	0.558	0.868
	Test	0.745	0.769	0.672–0.866	0.622	0.820	0.676	0.781
ResNet101	Training	0.901	0.939	0.908–0.971	0.875	0.919	0.875	0.919
	Validation	0.865	0.870	0.772–0.967	0.774	0.908	0.800	0.894
	Test	0.776	0.814	0.729–0.899	0.676	0.836	0.714	0.810
ResNet50	Training	0.771	0.873	0.828–0.918	0.830	0.733	0.670	0.868
	Validation	0.802	0.745	0.627–0.863	0.452	0.969	0.875	0.787
	Test	0.724	0.756	0.656–0.857	0.568	0.820	0.656	0.758
VGG19	Training	0.762	0.859	0.811–0.908	0.784	0.748	0.670	0.842
	Validation	0.812	0.852	0.758–0.945	0.742	0.846	0.697	0.873
	Test	0.806	0.773	0.664–0.883	0.649	0.902	0.800	0.809

AUC, area under curve; CI, confidence interval; DLR, deep learning radiomics; NPV, negative predictive value; PPV, positive predictive value.

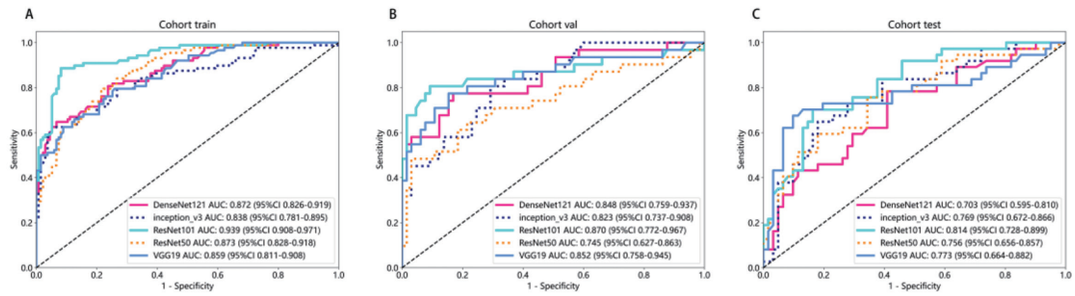

Figure S8 ROC results for deep learning model of different network.

Appendix 5: deep learning model

The ResNet101 network emerged as the superior performer, particularly highlighted by its AUC metric, making it the benchmark for comparison against other evaluated models.

ResNet101 showcased an impressive AUC of 0.939 in the training cohort, with a 95% CI of 0.9079–0.9706, indicating a strong predictive capability. In the validation cohort, it achieved an AUC of 0.870 (95% CI: 0.772–0.967). Furthermore, in the test cohort, ResNet101 maintained a commendable AUC of 0.814 (95% CI: 0.729–0.899). The details are shown in *Table S6* and *Figure S8*.

https://dx.doi.org/10.21037/tlcr-24-995